man lost in the convergence of time, Avebury (2022): Reconfiguring film through human figure removal and collage

Luís Arandas[1], Kate McDonough[2], Mick Grierson[3] and Miguel Carvalhais[4]

[1] University of Porto – INESC-TEC, Porto, Portugal
luis.arandas@inesctec.pt
[2] Anglia Ruskin University – The Otolith Group, Cambridge, United Kingdom
km9105@aru.ac.uk
[3] University of the Arts London – CCI, London, United Kingdom
m.grierson@arts.ac.uk
[4] i2ADS and Faculty of Fine Arts, University of Porto, Portugal
mcarvalhais@fba.up.pt

Abstract. This article outlines the processes taken in composing the piece *man lost in the convergence of time*, a series of prints in which generative diffusion models are used to abstract video through language, by initially removing the human from a film shoot at Avebury's ceremonial path, Wiltshire (2022). With this work we seek to reveal intrinsic bias in high-resolution image models released to the public for appropriation, introducing sets of *text prompts* which anchor *video-to-video* translation allegorical to science fiction. We developed a collage methodology where a secondary model opposed to the diffusion process predicts masks using dichotomous image segmentation, allowing us to composite the footage and mask to further diffusion steps recursively. Our series illustrates different attempts where we were successful in creating formal abstraction with semantic consistency and carry on a discussion on possible futures of experimental compositing techniques with CLIP-guidance conditioned by field recordings. We acknowledge our work promotes reinterpretation of space in moving image, as it has been in different generative model contexts.

Keywords: Alternative realities, language-guided diffusion, still-image composition, automatic compositing

1 Language-guided diffusion

Deep generative models have been proved very efficient at generating data according to a learned representation [5]. After learning procedures compute new data as a possible distribution sample and in themselves hold a simulation of multiple dimensions, representing found features [8]. Image diffusion appears as a successor of other generative models able to produce new image frames by synthesis according to natural language text [10]. Models which translate *text*

prompts into new image data promote what in the AI community is understood as *multimodality* [13], often with two different model architectures which interoperate via *embeddings* [14]. This research targets the condition of frame-by-frame diffusion as a methodology of abstraction; see [15] for contrasting still-image techniques.

2 The human figure as visual condition

The human figure has been paramount in image work using generative models, namely adversarial networks [3]. As a focus of both composition and production, human body is somehow represented in large datasets to train vision models [16]. Image diffusion models are able to compute sets of frames according to sets of text-prompts through embedding and coordination, see [11], and this allows practitioners to adapt personal longform manuscripts and generate new short films having natural language define how the picture should look like and how it develops in time; see Harun Farocki on the *construction of worlds* [2]. Using models to produce moving image is to leverage a computer system as a synthesizer of new data according to a learned reference, being these models many times rough simulations of the human brain's structure and behavior [7]. We draw on their ability to find similarities and reference different aspects of the world as scientific representation, even if they exist by their failure in practice [6]. Through training, independently and with divergent data types, minimise a loss and fit until a representation is accurate; with our work, we explore multimodality as a limit to be defined by the feedback of models interacting together: both the *language-image* mapping through CLIP-guidance but also adding a secondary dichotomous segmentation procedure [12] every frame of a produced film shoot. We removed the main human figure from each shot to then develop a compositing technique, embracing added failure on a topic which proved to be biased in the past[1]. To illustrate a simple CLIP-guided diffusion procedure conditioned by estimated masks we provide three examples (see Fig. 1); the extraction from the original frame with a minor diffusion for diagram interpretability (perceptually adding color not found through camera) and a final merge with the original.

[1] See *Salaf* (2020) by Nouf Aljowaysir, where the author tries to erase colonial echoes of classification in AI systems using this rationale, creating an *absent* dataset. We build on this removal as to cancel the human from the image, in our case to then diverge into new images by enforcing language descriptions.

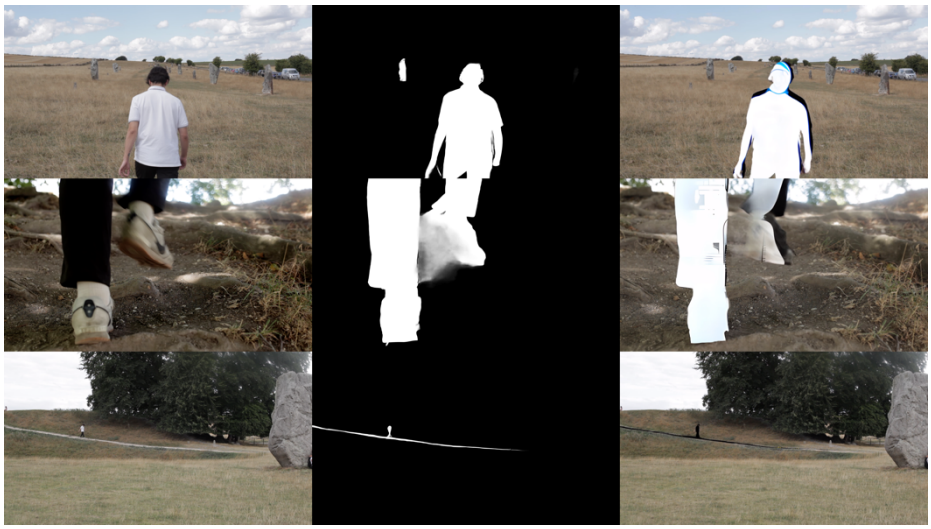

Fig. 1. Three compositing examples with predicted masks on three different locations.

3 Procedures and frame-set compositing

We provide an implementation[2] which one can build upon and composite their own datasets to condition new diffusion procedures, working with a torch compute graph using open implementations of both segmentation and CLIP-guided diffusion [10, 17]. Given an input image $\mathrm{I}(x, y)$, a binary mask $\mathrm{M}(x, y)$ representing the human figure, and the output image $\mathrm{O}(x, y)$[3], the binary mask $\mathrm{M}(x, y)$ is obtained algorithmically. At 25 FPS (frames per second), the equation has to be applied every 1/25 = 0.04 second (40 milliseconds) to create real-time video output:

$$O(x, y) = I(x, y) \cdot (1 - M(x, y)) + C \cdot M(x, y)$$

To understand this procedure as initial in the actual diffusion we propose: the output image $\mathrm{O}(x, y)$ from the compositing process be the initial frame $\mathrm{I}'(x, y)$, where the goal is to generate a final image $\mathrm{F}(x, y)$ that matches the desired properties specified by a textual prompt **P**. Each diffusion step generates a series of intermediate images $\mathrm{I}'(x, y, t)$ where $\boldsymbol{t}$ represents the diffusion timestep. If the generation process is guided by minimising the loss function **L**, which measures the discrepancy between the generated image's features and the desired features specified by the current text prompt **P**, the loss function can be defined as:

[2] https://github.com/luisArandas/guided-diffusion-segm-collage.

[3] Being that **x,y** are pixel coordinates and let **C** be the color of the human figure in the output image (white in this case, so **C** = 255 for an 8-bit grayscale).

$$L(I'(x,y,t),P) \;=\; D(C(I'(x,y,t)),C(P))$$

Here, $\mathrm{D}(a,b)$ represents a distance metric between two feature vectors **a** and **b**, and C(I) and C(P) represent the image and text prompt feature vectors, respectively, obtained from the CLIP model embedding. The CLIP-guided diffusion process involves iteratively updating $\mathrm{I}'(x,y,t)$ to minimise **L**, ultimately yielding the final image $\mathrm{F}(x,y)$ when the process converges. We denote the high-level description of our process as example that doesn't cover all the mechanisms inherent to actual implementation, e.g. optimising the loss, see [4].

4 Revealing through translation, a contemplative realm that no one really sees

We produced a series of prints *man lost in the convergence of time* using the mentioned procedures over a shoot in Avebury (Wiltshire, 2022)[4], contemplating the idea of morphological transformation through the help of language; see [18] on *video-to-video* synthesis. By acting on generative models we rely many times on sampling compressed representations with multiple dimensions, which themselves have been appropriated in production by defining trajectories [1]. Dimensions which map parts of records from the real world and allow us to introduce simulations and create connections between features and concepts that didn't exist before in practice. With this work we comply with the fact *text prompts* demand *to query* apart from previously proposed automatic content production research with the same diffusion process. By submitting ourselves to a conversational way of demanding how the film shoot could diverge aesthetically, we search for possible worlds of representation generative diffusion techniques can reveal, by demanding a fictional world of made-up portals and impossible visual objects after removing what might in fact be the fundamental link with the unknown. Practically, by not specifying what to detect when adding a secondary vision model, if successful the stones carved in the ground happen to be also removed. We believe there are poetic languages to be developed in time-based multimodal inference where by producing new methodologies of collage in the generative process, outputs might help us better understand decision and bias in moving image, reaffirming that explanations will always stay as not the reason of failure itself.

Acknowledgements

The research leading to these results and artwork was conducted at the UAL Creative Computing Institute (03-08/2022) and financially supported by the

[4] Part of a set of Neolithic and Bronze age ceremonial sites that seemingly formed a vast sacred landscape, south west England region. 9. Holgate, R., *Neolithic settlement patterns at Avebury, Wiltshire.* Antiquity, 1987. **61**(232): p. 259-263.

Portuguese Foundation for Science and Technology (FCT), through the individual research grant 2020.07619.BD and by the project "Experimentation in music in Portuguese culture: History, contexts and practices in the 20th and 21st centuries" (POCI-01-0145- FEDER-031380), co-funded by the European Union through the Operational Program Competitiveness and Internationalisation, in its ERDF component, and by national funds, through the Portuguese FCT.

References

1. Akten, M., R. Fiebrink, and M. Grierson, *Deep Meditations: Controlled navigation of latent space.* arXiv:2003.00910, 2020.
2. Almeida, J., P. Moran, and P. Arantes, *Harun Farocki: Programming the visible.* 2017.
3. Broad, T., F. Leymarie, and M. Grierson, *Amplifying The Uncanny.* 2020.
4. Cho, J., A. Zala, and M. Bansal, *Dall-eval: Probing the reasoning skills and social biases of text-to-image generative transformers.* arXiv preprint arXiv:2202.04053, 2022.
5. Croitoru, F.-A., et al., *Diffusion models in vision: A survey.* arXiv preprint arXiv:2209.04747, 2022.
6. Giere, R.N., *How models are used to represent reality.* Philosophy of science, 2004. **71**(5): p. 742-752.
7. Haenlein, M. and A. Kaplan, *A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence.* 2019.
8. Harshvardhan, G., et al., *A comprehensive survey and analysis of generative models in machine learning.* Computer Science Review, 2020. **38**: p. 100285.
9. Holgate, R., *Neolithic settlement patterns at Avebury, Wiltshire.* Antiquity, 1987. **61**(232): p. 259-263.
10. Kim, G., T. Kwon, and J.C. Ye. *DiffusionCLIP: Text-Guided Diffusion Models for Robust Image Manipulation*. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.
11. Muennighoff, N., *Sgpt: Gpt sentence embeddings for semantic search.* arXiv preprint arXiv:2202.08904, 2022.
12. Qin, X., et al., *Highly Accurate Dichotomous Image Segmentation.* 2022.
13. Radford, A., et al., *Learning Transferable Visual Models From Natural Language Supervision.* 2021.
14. Radford, A., et al., *Improving language understanding by generative pre-training.* 2018.
15. Ravi, H., et al., *PRedItOR: Text Guided Image Editing with Diffusion Prior.* arXiv preprint arXiv:2302.07979, 2023.
16. Ridnik, T., et al., *Imagenet-21k pretraining for the masses.* arXiv preprint arXiv:2104.10972, 2021.
17. Subramanian, V., *Deep Learning with PyTorch: A practical approach to building neural network models using PyTorch*. 2018: Packt Publishing Ltd.
18. Wang, T.-C., et al., *Video-to-video synthesis.* arXiv preprint arXiv:1808.06601, 2018.